

Introduction to Big Data Assignment-1

Name: Faizan Mulla

Roll No: 21F1003885

Problem Statement

Write a Spark code to analyze user click patterns throughout the day by categorizing clicks into different time intervals (0-6, 6-12, 12-18, and 18-24 hours) using data stored in Google Cloud Storage.

Solution

I approached the problem by implementing a two-part solution using Google Cloud Platform (GCP). The solution consists of:

1. Data generator script to create csv file
2. Spark analysis script to process the data and generate required solution

Environment Setup

1. GCP Setup

- Set up a new project with default service account configurations
- Enabled necessary APIs (Compute Engine and Dataproc)

2. Dataproc Cluster Creation

- Created a Dataproc cluster with following specifications:
 - chose the option: Create cluster on compute engine
 - manager node: series → e2 // machine type → e2-standard-2 (2vCPU, 1 core, 8GB memory)
 - reduce primary disk size from 500GB to something less like 50GB.
 - exact same settings for worker nodes too.

- Region: **asia-south-1**
- in the customize cluster menu, **uncheck** the INTERNAL IP ONLY option.

3. Cloud Storage Setup

- Created a Cloud Storage bucket for storing input and output files
- Configured with standard storage class
- Region: **asia-south-1**

Implementation Details

1. Data Generation Script

```
import csv
import random
from datetime import timedelta, datetime

def generate_random_timestamp(base_date=datetime(2024, 10, 15)):
    return base_date + timedelta(
        days=random.randint(0, 365),
        hours=random.randint(0, 23),
        minutes=random.randint(0, 59),
        seconds=random.randint(0, 59),
    )

data = {f"user_{i + 1}": generate_random_timestamp() for i in range(0, 50)}

output_file = "data.csv"

with open(output_file, "w", newline="") as file:
    writer = csv.writer(file)
    writer.writerow(["id", "timestamp", "date"])
    writer.writerows([
        (user_id, ts.strftime("%Y-%m-%d %H:%M:%S"), ts.strftime("%Y-%m-%d"))
        for user_id, ts in data.items()
    ])
]
```

2. Spark Analysis Script

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, hour, when

spark = SparkSession.builder.appName("user_click_counter").getOrCreate()
input_file = "gs://ibd-ga3/data.csv"
output_file = "output.txt"

data = spark.read.option("header", "true").option("delimiter", ",").csv(input_file)

data = data.withColumn("Hour", hour(data["timestamp"]))

data = data.withColumn(
    "time_interval",
    when(col("Hour") < 6, "00-06")
    .when(col("Hour") < 12, "06-12")
    .when(col("Hour") < 18, "12-18")
    .when(col("Hour") < 24, "18-24")
    .otherwise("Invalid timestamp"),
)

data = data.sort("time_interval")
result = data.groupBy("time_interval").count().sort("time_interval")
result.show()
result.toPandas().to_csv(output_file, sep="\t", index=False, header=True)

spark.stop()
```

Execution Process

1. Data Generation

- Uploaded **data-generator.py** to Cloud Storage
- Executed script to create sample dataset
- Verified **data.csv** in Cloud Storage

2. Spark Analysis

- Uploaded **spark.py** to Cloud Storage
- Submitted Spark job through Dataproc
- Monitored job execution through Dataproc UI

Results

The analysis successfully categorized user clicks into four time intervals:

- 00-06: **14** clicks
- 06-12: **13** clicks
- 12-18: **10** clicks
- 18-24: **13** clicks

Relevant Screenshots

1. Dataproc Cluster Configuration

The screenshot displays the Google Cloud Dataproc Clusters page. The top navigation bar includes the Google Cloud logo, a project selector (My First Project), and a search bar. The left sidebar shows navigation options: Clusters, Jobs, Workflows, Autoscaling policies, and Serverless. The main content area shows a table of clusters with columns: Name, Status, Region, Zone, Total worker nodes, Flexible VMs?, Scheduled deletion, Cloud Storage staging bucket, Created, and Labels. A cluster named 'litm-lbd-ga3' is shown in a 'Running' state. Below the table, a detailed view of the cluster is shown, including a warning message about permissions and a table of cluster details.

Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging bucket	Created	Labels
litm-lbd-ga3	Running	asia-south1	asia-south1-a	2	No	Off	dataproc-staging-asia-south1-722420454652-nlwmyff	Oct 20, 2024, 7:29:49 PM	goog-dataproc-enabled

Warning: Failed to validate permissions required for default service account: '722420454652-compute@developer.gserviceaccount.com'. Cluster creation could still be successful if required permissions have been granted to the respective service accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project '722420454652' before or it is disabled. Enable it by visiting <https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=722420454652>.

Name	litm-lbd-ga3
Cluster UUID	9441ec75-0a54-43e2-8934-831b4292135a
Type	Dataproc Cluster
Status	Running

2. Cloud Storage Bucket Contents

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

Cloud Storage

Bucket details

GO TO PATH REFRESH LEARN

Overview

Buckets

Monitoring

Settings

ibd-ga3

Location: us (multiple regions in United States)

Storage class: Standard

Public access: Not public

Protection: Soft Delete

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

OPERATIONS

Folder browser

ibd-ga3

CREATE FOLDER UPLOAD TRANSFER DATA OTHER SERVICES

Filter by name prefix only Filter objects and folders

Show Live objects only

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	
data.csv	2 KB	text/csv	Oct 20, 2024, 7:27:22 PM	Standard	Oct 20, 2024, 7:27:22 PM	Not public	—	+
spark.py	830 B	text/x-python	Oct 20, 2024, 7:28:29 PM	Standard	Oct 20, 2024, 7:28:29 PM	Not public	—	+

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

Cloud Storage

Object details

LEARN

Overview

Buckets

Monitoring

Settings

Buckets > ibd-ga3 > spark.py

LIVE OBJECT VERSION HISTORY

DOWNLOAD EDIT METADATA EDIT ACCESS DELETE

Overview

Type	text/x-python
Size	830 B
Created	Oct 20, 2024, 7:28:29 PM
Last modified	Oct 20, 2024, 7:28:29 PM
Storage class	Standard
Custom time	—
Public URL	Not applicable
Authenticated URL	https://storage.cloud.google.com/ibd-ga3/spark.py
gsutil URI	gs://ibd-ga3/spark.py

Permissions

Public access	Not public
---------------	------------

Protection

Version history	—
Retention expiration time	None
Object retention retain until time	None
Bucket retention retain until time	None
Hold status	None
Encryption type	Google-managed

Marketplace

Release Notes

3. Job Execution Results

Job details

CLONE DELETE STOP REFRESH

Job ID	iitm-ibd-ga3-job
Job UUID	1902d7f0-f4b1-4183-9663-e5c051ac4071
Type	Dataproc Job
Status	Succeeded

MONITORING

CONFIGURATION

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

Output

LINE WRAP: OFF

Spark jobs take ~60 seconds to initialize resources.

```

+-----+
|time_interval|count|
+-----+
| 00-06 | 14 |
| 06-12 | 13 |
| 12-18 | 10 |
| 18-24 | 13 |
+-----+
```

4. Output File Contents

Free trial status: ₹25,092.75 credit and 90 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

DISMISS

ACTIVATE

Google Cloud

My First Project

Search (/) for resources, docs, products, and more

Search

Dataproc

Job details

CLONE

DELETE

STOP

REFRESH

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Interactive

Interactive Templates

Metastore Services

Metastore

Federation

Release Notes

Output

LINE WRAP: OFF

DISMISS

Spark jobs take ~60 seconds to initialize resources.

24/10/20 14:15:53 INFO MetricsSystemImpl: google-hadoop-file-system metrics system started
24/10/20 14:15:54 INFO DataprocSparkPlugin: Registered 171 driver metrics
24/10/20 14:15:56 INFO DefaultHARMAccessProxyProvider: Connecting to ResourceManager at iitm-ibd-ga3-n.asia-south1-a.c.starlit-link-439117-v6.internal./10.160.0.7:8032
24/10/20 14:15:56 INFO AHSProxy: Connecting to Application History server at iitm-ibd-ga3-n.asia-south1-a.c.starlit-link-439117-v6.internal./10.160.0.7:10200
24/10/20 14:15:58 INFO Configuration: resource-types.xml not found
24/10/20 14:15:58 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/10/20 14:16:00 INFO YarnClientImpl: Submitted application application_1729432572597_0001
24/10/20 14:16:00 INFO DefaultHARMAccessProxyProvider: Connecting to ResourceManager at iitm-ibd-ga3-n.asia-south1-a.c.starlit-link-439117-v6.internal./10.160.0.7:8032
24/10/20 14:16:05 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-asia-south1-722420454652-puzokboi/objects?fields=item:
24/10/20 14:16:05 INFO GhsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_error_count=1, gcs_api_time=1255, gcs_api_t
24/10/20 14:16:06 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/10/20 14:16:09 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2ops]); readers will "met" yet see flushed data for gs://dataproc-temp-asia-sou
24/10/20 14:16:15 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/ibd-ga2/o/data.csv?fields=bucket,name,timeCreated,updated,generation,metad
24/10/20 14:16:27 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-asia-south1-722420454652-puzokboi/o/9441ec75-0a54-43e2-8934-
24/10/20 14:16:42 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-asia-south1-722420454652-puzokboi/o/9441ec75-0a54-43e2-8934-
+-----+
|time_interval|count|
+-----+
00-06	14
06-12	13
12-18	10
18-24	13
+-----+	
24/10/20 14:16:54 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics={action_http_patch_request=0, files_created=1, gcs_api_server_timeout_count=0, op_get_list_status_result_si	
24/10/20 14:16:54 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dataproc-temp-asia-south1-722420454652-puzokboi/o/9441ec75-0a54-43e2-8934-	
+-----+	
time_interval	count
+-----+	
00-06	14
06-12	13
12-18	10
18-24	13
+-----+